Translation

PATENT COOPERATION TREATY

PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

7

Applicant's or agent's file reference EM99102_WO	FOR FURTHER ACTION	See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416)
International application No. PCT/EP00/07953	International filing date (*day/month/year*) 16 August 2000 (16.08.00)	Priority date (*day/month/year*) 01 September 1999 (01.09.99)
International Patent Classification (IPC) or national classification and IPC C12Q 1/68,		
Applicant MERCK PATENT GMBH		

1. This international preliminary examination report has been prepared by this International Preliminary Examining Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 8 sheets, including this cover sheet.

☒ This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

These annexes consist of a total of 3 sheets.

3. This report contains indications relating to the following items:

- I ☒ Basis of the report
- II ☐ Priority
- III ☐ Non-establishment of opinion with regard to novelty, inventive step and industrial applicability
- IV ☐ Lack of unity of invention
- V ☒ Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement
- VI ☒ Certain documents cited
- VII ☒ Certain defects in the international application
- VIII ☒ Certain observations on the international application

Date of submission of the demand	Date of completion of this report
24 March 2001 (24.03.01)	05 December 2001 (05.12.2001)
Name and mailing address of the IPEA/EP Facsimile No.	Authorized officer Telephone No.

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/EP00/07953

I. Basis of the report

1. This report has been drawn on the basis of (*Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as "originally filed" and are not annexed to the report since they do not contain amendments.*):

☒ the international application as originally filed.

☒ the description, pages 1-11, as originally filed,
pages _____, filed with the demand,
pages _____, filed with the letter of _____.
pages _____, filed with the letter of _____.

☒ the claims, Nos. _____, as originally filed,
Nos. _____, as amended under Article 19,
Nos. _____, filed with the demand,
Nos. 1-14, filed with the letter of 26 November 2001 (26.11.2001).
Nos. _____, filed with the letter of _____.

☒ the drawings, sheets/fig 1/2, 2/2, as originally filed,
sheets/fig _____, filed with the demand,
sheets/fig _____, filed with the letter of _____,
sheets/fig _____, filed with the letter of _____.

2. The amendments have resulted in the cancellation of:

☐ the description, pages _____
☐ the claims, Nos. _____
☐ the drawings, sheets/fig _____

3. ☐ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).

4. Additional observations, if necessary:

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty (N)	Claims	1-14	YES
	Claims		NO
Inventive step (IS)	Claims	1-14	YES
	Claims		NO
Industrial applicability (IA)	Claims	1-14	YES
	Claims		NO

2. Citations and explanations

This report makes reference to the following documents:

D1 Genome Research, Vol. 5, 1995, pages 173 to 184

D2 Methods in Enzymology, Vol. 266, 1996, pages 131 to 141

D3 Information available at "www.ncbi.nlm.nih.gov/UniGene" since August 1997

1. Claim 1 relates to a method for determining potentially significant DNA and/or nucleic acid sequences of a species of interest (species sequences) with the following steps:
a) determining any species sequences of species of interest using biological or genetically engineered methods and storing the species sequences in a first database,
b) detecting known DNA/ nucleic acid sequences of a predetermined group of different types (biosequences) including the functional significance of these sequences, in a second database, in which the biosequences and additional information including the functional significance of individual biosequences are stored,

c) comparing in a homology test the already known species sequences of the species of interest with the biosequences of the predetermined group of biosequences stored in the second database,
d) selecting those biosequences of the predetermined group that are homologous with the known species sequences over a predetermined threshold,
e) comparison in a second homologous test of the biosequences not selected and remaining in the second database from the group mentioned with the species sequences determined as described in step a)
f) storing and/or issuing those species sequences as species sequences of potentially increased importance, whose homology with biosequences made up of the group mentioned of remaining biosequences exceeds a predetermined second threshold value, along with information about the respective homologous biosequences
g) it being possible to carry out step e) optionally before step c) and without the preceding selection described in step d)
i) classification of the species sequences given and stored in step f), i.e. ordering (sorting) into particular classes of sequences by linguistic analysis of text definitions of the additional information stored about the homologous biosequences.

Such a method is known from the available prior art. It thus meets the requirements of PCT Article 33(2). The same remark applies to the subject matter of dependent Claims 2 to 14.

2. Moreover, such a method seems to involve an inventive step.

D1 is considered to be the closest prior art. This document discloses a method for automated sequence analysis by involving several databases. D1 describes the improved BLAST sequence alignment auxiliary agent "BEAUTY" which can help to identify not only homologous sequences already known for a newly sequenced DNA or protein section but also can help to request functional data. Functional data is extracted by providing "links" to other protein or DNA databases available on the Internet, such as Medline or OMIM (abstract, page 173, second column, second paragraph to page 174, second column, third paragraph; page 175, table 1, first column, first paragraph; page 180, Figure 5). The subject matter of Claim 1 differs from this prior art in that two homology tests are done. After the first test all DNA sequences that are above a predetermined homology threshold value are selected, thereby reducing the calculating effort when comparing the second sequence.

The problem addressed by the present application was, accordingly, how to develop a more rapid DNA sequence comparison method.

This aforementioned problem was solved by reducing the sequence data stock in the second database first by subtraction with already known data from the first database with the result that the second homology comparison requires less calculating capacity, resulting in accelerated sequence data alignment. This method step was not known from the available documents. Moreover, the resulting technical effect does not seem to be suggested by

the available documents. The application can therefore be said to involve an inventive step (PCT Article 33(3)). The same remark applies to the subject matter of Claims 2 to 14 that are dependent on this claim.

Supplemental Box
(To be used when the space in any of the preceding boxes is not sufficient)

Continuation of: VI

The search report citation WO-A-00 63687, filed on 14.04.2000, published on 26.10.2000 with the priority data of 15.04.1999 and WO-A-01 13105, filed on 28.07.2000, published on 22.01.2001 with the priority date of 30.07.1999, might be relevant for the subject matter of the present application should the claimed priority of the present claims not be valid.

VII. Certain defects in the international application

The following defects in the form or contents of the international application have been noted:

Contrary to PCT Rule 5.1(a)(ii), the description does not cite D1 and D2 or indicate the relevant prior art disclosed therein.

VIII. Certain observations on the international application

The following observations on the clarity of the claims, description, and drawings or on the question whether the claims are fully supported by the description, are made:

1. The following phrases used in Claims 1 and 2 and vague and not clear and leave the reader uncertain about the meaning of the relevant technical features.

Claim 1 (d): "are homologous over a predetermined threshold"

Claim 1 (f): "exceeds a predetermined threshold"

Claim 1 (i): "by a linguistic analysis of text definitions"

Claim 2 (h): "in adaptation, optimised according to predetermined criteria, to the respective homologous biosequences".

As a result the definition of the subject matter of these claims is not clear (PCT Article 6). More particularly, the technical criteria by means of which the homology threshold values are determined (80%, 90% 100% homology?) are not clear. In step 1(d), firstly "homologous" sequences are removed from the comparative database, whereas in the second homology test more "homologous" sequences are found. Moreover, the claimed scope and the subject matter of Claim 2 are not clear at all.

2. Moreover, the description does not give any technical teaching to indicate how the "computer program" of Claim 1 was programmed. Consequently, it seems to be disclosed in a form that cannot be carried out by a person skilled in the art (PCT Article 5).

VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS

PCT

INTERNATIONALER RECHERCHENBERICHT

(Artikel 18 sowie Regeln 43 und 44 PCT)

Aktenzeichen des Anmelders oder Anwalts	**WEITERES VORGEHEN**	siehe Mitteilung über die Übermittlung des internationalen Recherchenberichts (Formblatt PCT/ISA/220) sowie, soweit zutreffend, nachstehender Punkt 5
EM99102_WO		
Internationales Aktenzeichen	Internationales Anmeldedatum *(Tag/Monat/Jahr)*	(Frühestes) Prioritätsdatum *(Tag/Monat/Jahr)*
PCT/EP 00/ 07953	16/08/2000	01/09/1999
Anmelder		
MERCK PATENT GMBH		

Dieser internationale Recherchenbericht wurde von der Internationalen Recherchenbehörde erstellt und wird dem Anmelder gemäß Artikel 18 übermittelt. Eine Kopie wird dem Internationalen Büro übermittelt.

Dieser internationale Recherchenbericht umfaßt insgesamt 4 Blätter.

[X] Darüber hinaus liegt ihm jeweils eine Kopie der in diesem Bericht genannten Unterlagen zum Stand der Technik bei.

1. **Grundlage des Berichts**

   a. Hinsichtlich der **Sprache** ist die internationale Recherche auf der Grundlage der internationalen Anmeldung in der Sprache durchgeführt worden, in der sie eingereicht wurde, sofern unter diesem Punkt nichts anderes angegeben ist.

   [ ] Die internationale Recherche ist auf der Grundlage einer bei der Behörde eingereichten Übersetzung der internationalen Anmeldung (Regel 23.1 b)) durchgeführt worden.

   b. Hinsichtlich der in der internationalen Anmeldung offenbarten **Nucleotid- und/oder Aminosäuresequenz** ist die internationale Recherche auf der Grundlage des Sequenzprotokolls durchgeführt worden, das

   [ ] in der internationalen Anmeldung in schriftlicher Form enthalten ist.

   [ ] zusammen mit der internationalen Anmeldung in computerlesbarer Form eingereicht worden ist.

   [ ] bei der Behörde nachträglich in schriftlicher Form eingereicht worden ist.

   [ ] bei der Behörde nachträglich in computerlesbarer Form eingereicht worden ist.

   [ ] Die Erklärung, daß das nachträglich eingereichte schriftliche Sequenzprotokoll nicht über den Offenbarungsgehalt der internationalen Anmeldung im Anmeldezeitpunkt hinausgeht, wurde vorgelegt.

   [ ] Die Erklärung, daß die in computerlesbarer Form erfaßten Informationen dem schriftlichen Sequenzprotokoll entsprechen, wurde vorgelegt.

2. [ ] **Bestimmte Ansprüche haben sich als nicht recherchierbar erwiesen** (siehe Feld I).

3. [ ] **Mangelnde Einheitlichkeit der Erfindung** (siehe Feld II).

4. Hinsichtlich der **Bezeichnung der Erfindung**

   [X] wird der vom Anmelder eingereichte Wortlaut genehmigt.

   [ ] wurde der Wortlaut von der Behörde wie folgt festgesetzt:

5. Hinsichtlich der **Zusammenfassung**

   [ ] wird der vom Anmelder eingereichte Wortlaut genehmigt.

   [X] wurde der Wortlaut nach Regel 38.2b) in der in Feld III angegebenen Fassung von der Behörde festgesetzt. Der Anmelder kann der Behörde innerhalb eines Monats nach dem Datum der Absendung dieses internationalen Recherchenberichts eine Stellungnahme vorlegen.

6. Folgende Abbildung der **Zeichnungen** ist mit der Zusammenfassung zu veröffentlichen: Abb. Nr. 1

   [X] wie vom Anmelder vorgeschlagen

   [ ] keine der Abb.

   [ ] weil der Anmelder selbst keine Abbildung vorgeschlagen hat.

   [ ] weil diese Abbildung die Erfindung besser kennzeichnet.

Feld III WORTLAUT DER ZUSAMMENFASSUNG (Fortsetzung von Punkt 5 auf Blatt 1)

Die vorliegende Erfindung betrifft ein Verfahren zum Ermitteln potentiell bedeutsamer DNA- und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen). Um ein Verfahren zum Ermitteln von DNA- und/oder Nukleinsäuresequenzen zu schaffen, bei welchem gezielt solche DNA- und/oder Nukleinsäuresequenzen herausselektiert werden, die eine potentiell erhöhte Bedeutsamkeit haben, das heißt die mit erheblich weniger Forschungsaufwand gezielt im Hinblick auf bestimmte Funktionen untersucht werden können, insbesondere im Hinblick auf eine potentielle Kranheitsrelevanz.

Internationales Aktenzeichen
P[redacted] 00/07953

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
Y	WORLEY K C ET AL: "BEAUTY: AN ENHANCED BLAST-BASED SEARCH TOOL THAT INTEGRATES MULTIPLE BIOLOGICAL INFORMATION RESOURCES INTO SEQUENCE SIMILARITY SEARCH RESULTS" GENOME RESEARCH,US,COLD SPRING HARBOR LABORATORY PRESS, Bd. 5, Nr. 2, 1. September 1995 (1995-09-01), Seiten 173-184, XP000534406 ISSN: 1088-9051 das ganze Dokument	1-15
E	WO 01 13105 A (CHIN DANIEL J ;HENDRIX DONNA (US); ZHAO OLIVER (US); AGY THERAPEUT) 22. Februar 2001 (2001-02-22) Zusammenfassung; Ansprüche 1-13	1-15
E	WO 00 63687 A (UNIV COLUMBIA) 26. Oktober 2000 (2000-10-26) Zusammenfassung; Anspruch 1 Seite 44, Zeile 5 -Seite 45, Zeile 10	1-15
A	US 5 871 697 A (DEEM MICHAEL W ET AL) 16. Februar 1999 (1999-02-16) Zusammenfassung; Ansprüche 1-6 Spalte 58, Absatz 2 -Spalte 59, Absatz 2	1-15

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP 00/07953

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0113105 A	22-02-2001	AU 6611900 A	13-03-2001
WO 0063687 A	26-10-2000	AU 4355600 A	02-11-2000
US 5871697 A	16-02-1999	AU 730830 B	15-03-2001
		AU 7476396 A	15-05-1997
		EP 0866877 A	30-09-1998
		JP 2000500647 T	25-01-2000
		WO 9715690 A	01-05-1997
		US 6231812 B	15-05-2001
		US 5972693 A	26-10-1999
		US 2001007985 A	12-07-2001
		US 6141657 A	31-10-2000

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PC[redacted] 00/07953

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 C12Q1/68 G06F19/00

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06F

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, WPI Data

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
Y	MADDEN T L ET AL: "APPLICATIONS OF NETWORK BLAST SERVER" METHODS IN ENZYMOLOGY,ACADEMIC PRESS INC, SAN DIEGO, CA,US, Bd. 266, 1996, Seiten 131-141, XP001006313 ISSN: 0076-6879 das ganze Dokument --- -/--	1-15

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

° *Besondere Kategorien von angegebenen Veröffentlichungen* :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" *Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft er-*scheinen zu lassen, oder durch die das Veröffentlichungsdatum einer *anderen im Recherchenbericht genannten Veröffentlichung belegt werden* soll oder die aus einem anderen besonderen Grund angegeben ist (wie *ausgeführt*)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung *kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf* erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
2. August 2001	09/08/2001

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Filloy García, E

1

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION

(PCT Rule 61.2)

To:

Commissioner
US Department of Commerce
United States Patent and Trademark Office, PCT
2011 South Clark Place Room
CP2/5C24
Arlington, VA 22202
ETATS-UNIS D'AMERIQUE

in its capacity as elected Office

Date of mailing (day/month/year) 11 May 2001 (11.05.01)	
International application No. PCT/EP00/07953	**Applicant's or agent's file reference** EM99102_WO
International filing date (day/month/year) 16 August 2000 (16.08.00)	**Priority date** (day/month/year) 01 September 1999 (01.09.99)
Applicant TOLDO, Luca et al	

1. The designated Office is hereby notified of its election made:

[X] in the demand filed with the International Preliminary Examining Authority on:

24 March 2001 (24.03.01)

[ ] in a notice effecting later election filed with the International Bureau on:

2. The election [X] was

[ ] was not

made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland**	Authorized officer Juan Cruz
Facsimile No.: (41-22) 740.14.35	Telephone No.: (41-22) 338.83.38

VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS

10 DEC 2001

PCT

INTERNATIONALER VORLÄUFIGER PRÜFUNGSBERICHT

(Artikel 36 und Regel 70 PCT)

Aktenzeichen des Anmelders oder Anwalts E/M99102_WO	**WEITERES VORGEHEN** siehe Mitteilung über die Übersendung des internationalen vorläufigen Prüfungsberichts (Formblatt PCT/IPEA/416)	
Internationales Aktenzeichen PCT/EP00/07953	Internationales Anmeldedatum *(Tag/Monat/Jahr)* 16/08/2000	Prioritätsdatum *(Tag/Monat/Tag)* 01/09/1999
Internationale Patentklassifikation (IPK) oder nationale Klassifikation und IPK C12Q1/68		
Anmelder MERCK PATENT GMBH et al.		

1. Dieser internationale vorläufige Prüfungsbericht wurde von der mit der internationalen vorläufigen Prüfung beauftragten Behörde erstellt und wird dem Anmelder gemäß Artikel 36 übermittelt.

2. Dieser BERICHT umfaßt insgesamt 8 Blätter einschließlich dieses Deckblatts.

   ☒ Außerdem liegen dem Bericht ANLAGEN bei; dabei handelt es sich um Blätter mit Beschreibungen, Ansprüchen und/oder Zeichnungen, die geändert wurden und diesem Bericht zugrunde liegen, und/oder Blätter mit vor dieser Behörde vorgenommenen Berichtigungen (siehe Regel 70.16 und Abschnitt 607 der Verwaltungsrichtlinien zum PCT).

   Diese Anlagen umfassen insgesamt 3 Blätter.

3. Dieser Bericht enthält Angaben zu folgenden Punkten:

   - I ☒ Grundlage des Berichts
   - II ☐ Priorität
   - III ☐ Keine Erstellung eines Gutachtens über Neuheit, erfinderische Tätigkeit und gewerbliche Anwendbarkeit
   - IV ☐ Mangelnde Einheitlichkeit der Erfindung
   - V ☒ Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung
   - VI ☒ Bestimmte angeführte Unterlagen
   - VII ☒ Bestimmte Mängel der internationalen Anmeldung
   - VIII ☒ Bestimmte Bemerkungen zur internationalen Anmeldung

Datum der Einreichung des Antrags	Datum der Fertigstellung dieses Berichts
24/03/2001	05.12.2001
Name und Postanschrift der mit der internationalen vorläufigen Prüfung beauftragten Behörde: Europäisches Patentamt D-80298 München Tel. +49 89 2399 - 0 Tx: 523656 epmu d Fax: +49 89 2399 - 4465	Bevollmächtigter Bediensteter Montrone, M Tel. Nr. +49 89 2399 8711

Formblatt PCT/IPEA/409 (Deckblatt) (Januar 1994)

I. Grundlage des Berichts

1. Hinsichtlich der **Bestandteile** der internationalen Anmeldung (*Ersatzblätter, die dem Anmeldeamt auf eine Aufforderung nach Artikel 14 hin vorgelegt wurden, gelten im Rahmen dieses Berichts als "ursprünglich eingereicht" und sind ihm nicht beigefügt, weil sie keine Änderungen enthalten (Regeln 70.16 und 70.17)*): **Beschreibung, Seiten:**

   1-11 ursprüngliche Fassung

   Patentansprüche, Nr.:

   1-14 mit Telefax vom 26/11/2001

   Zeichnungen, Blätter:

   1/2,2/2 ursprüngliche Fassung

2. Hinsichtlich der **Sprache**: Alle vorstehend genannten Bestandteile standen der Behörde in der Sprache, in der die internationale Anmeldung eingereicht worden ist, zur Verfügung oder wurden in dieser eingereicht, sofern unter diesem Punkt nichts anderes angegeben ist.

   Die Bestandteile standen der Behörde in der Sprache: zur Verfügung bzw. wurden in dieser Sprache eingereicht; dabei handelt es sich um

   ☐ die Sprache der Übersetzung, die für die Zwecke der internationalen Recherche eingereicht worden ist (nach Regel 23.1(b)).

   ☐ die Veröffentlichungssprache der internationalen Anmeldung (nach Regel 48.3(b)).

   ☐ die Sprache der Übersetzung, die für die Zwecke der internationalen vorläufigen Prüfung eingereicht worden ist (nach Regel 55.2 und/oder 55.3).

3. Hinsichtlich der in der internationalen Anmeldung offenbarten **Nucleotid- und/oder Aminosäuresequenz** ist die internationale vorläufige Prüfung auf der Grundlage des Sequenzprotokolls durchgeführt worden, das:

   ☐ in der internationalen Anmeldung in schriftlicher Form enthalten ist.

   ☐ zusammen mit der internationalen Anmeldung in computerlesbarer Form eingereicht worden ist.

   ☐ bei der Behörde nachträglich in schriftlicher Form eingereicht worden ist.

   ☐ bei der Behörde nachträglich in computerlesbarer Form eingereicht worden ist.

   ☐ Die Erklärung, daß das nachträglich eingereichte schriftliche Sequenzprotokoll nicht über den Offenbarungsgehalt der internationalen Anmeldung im Anmeldezeitpunkt hinausgeht, wurde vorgelegt.

   ☐ Die Erklärung, daß die in computerlesbarer Form erfassten Informationen dem schriftlichen Sequenzprotokoll entsprechen, wurde vorgelegt.

4. Aufgrund der Änderungen sind folgende Unterlagen fortgefallen:

- ☐ Beschreibung, Seiten:
- ☐ Ansprüche, Nr.:
- ☐ Zeichnungen, Blatt:

5. ☐ Dieser Bericht ist ohne Berücksichtigung (von einigen) der Änderungen erstellt worden, da diese aus den angegebenen Gründen nach Auffassung der Behörde über den Offenbarungsgehalt in der ursprünglich eingereichten Fassung hinausgehen (Regel 70.2(c)).

   (Auf Ersatzblätter, die solche Änderungen enthalten, ist unter Punkt 1 hinzuweisen;sie sind diesem Bericht beizufügen).

6. Etwaige zusätzliche Bemerkungen:

V. Begründete Feststellung nach Artikel 35(2) hinsichtlich der Neuheit, der erfinderischen Tätigkeit und der gewerblichen Anwendbarkeit; Unterlagen und Erklärungen zur Stützung dieser Feststellung

1. Feststellung

Neuheit (N)	Ja:	Ansprüche	1-14
	Nein:	Ansprüche	
Erfinderische Tätigkeit (ET)	Ja:	Ansprüche	1-14
	Nein:	Ansprüche	
Gewerbliche Anwendbarkeit (GA)	Ja:	Ansprüche	1-14
	Nein:	Ansprüche	

2. Unterlagen und Erklärungen
   siehe Beiblatt

VI. Bestimmte angeführte Unterlagen

1. Bestimmte veröffentlichte Unterlagen (Regel 70.10)

und / oder

2. Nicht-schriftliche Offenbarungen (Regel 70.9)

siehe Beiblatt

VII. Bestimmte Mängel der internationalen Anmeldung

Es wurde festgestellt, daß die internationale Anmeldung nach Form oder Inhalt folgende Mängel aufweist:
siehe Beiblatt

VIII. Bestimmte Bemerkungen zur internationalen Anmeldung

Zur Klarheit der Patentansprüche, der Beschreibung und der Zeichnungen oder zu der Frage, ob die Ansprüche in vollem Umfang durch die Beschreibung gestützt werden, ist folgendes zu bemerken:
siehe Beiblatt

Es wird auf die folgenden Dokumente verwiesen:

D1: Genome Research, Bd. 5, 1995, Seiten 173-184
D2: Methods in Enzymology, Bd. 266, 1996, Seiten 131-141
D3: Informationen unter "www.ncbi.nlm.nih.gov/UniGene" verfügbar seit August 1997.

Punkt V:

1. Anspruch 1 bezieht sich auf ein Verfahren zum Ermitteln potentiell bedeutsamer DNA-und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen) mit den folgenden Schritten :
a) Ermitteln beliebiger Artsequenzen der interessierenden Spezies mit biologischen bzw. gentechnischen Methoden und Speichern der Artsequenzen in einer ersten Datenbank,
b) Erfassen bekannter DNA-/Nukleinsäuresequenzen einer vorgegebenen Gruppe anderer Arten (Biosequenzen) einschließlich der funktionalen Bedeutung dieser Sequenzen, in einer zweiten Datenbank, in welcher die Biosequenzen und Zusatzinformationen einschließlich der funktionalen Bedeutung einzelner Biosequenzen gespeichert sind,
c) Vergleichen der bereits bekannten Artsequenzen der interessierenden Spezies mit den Biosequenzen der in der zweiten Datenbank gespeicherten, vorgegebenen Gruppe von Biosequenzen in einem Homologietest,
d) Aussondern derjenigen Biosequenzen der vorgegebenen Gruppe, die zu den bekannten Artsequenzen über einem vorgegebenen Schwellenwert homolog sind,
e) Vergleichen der aus der zweiten Datenbank verbleibenden, nicht ausgesonderten Biosequenzen aus der erwähnten Gruppe mit den nach Schritt a ermittelten Artsequenzen in einem zweiten Homologietest,
f) Speichern und/oder Ausgeben derjenigen Artsequenzen als Artsequenzen potentiell erhöhter Bedeutung, deren Homologie mit Biosequenzen aus der

erwähnten Gruppe verbliebenen Biosequenzen einen vorgegebenen zweiten Schwellenwert überschreitet, zusammen mit Informationen über die hierzu jeweils homologen Biosequenzen.
g) Wobei Schritt e) wahlweise auch vor Schritt c) und ohne vorheriges Aussondern gemäß Schritt d) durchführbar ist
i) Klassifizieren der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen, d.h. Einordnen (Sortieren) in bestimmte Klassen von Sequenzen durch linguistische Analyse von Textdefinitionen der zu den homologen Biosequenzen gespeicherten Zusatzinformationen.

Ein solches Verfahren ist aus dem verfügbaren Stand der Technik nicht bekannt. Es erfüllt somit die Erfordernisse von Art. 33(2) PCT. Dasselbe gilt für den Gegenstand der davon abhängigen Ansprüche 2 bis 14.

2. Darüber hinaus erscheint ein solches Verfahren auf einer erfinderischen Tätigkeit zu beruhen.

D1 wird als nächster Stand der Technik erachtet. Dieses Dokument offenbart ein Verfahren zur automatisierten Sequenzanalyse unter Einbeziehung mehrerer Datenbanken. D1 beschreibt das verbesserte BLAST Sequenzalignment Hilfsmittel "BEAUTY" mit dessen Hilfe nicht nur homologe bereits bekannte Sequenzen für einen neu sequenzierten DNA- oder Proteinabschnitt identifiziert werden können, sondern gleichzeitig auch funktionelle Daten abgerufen werden können. Die Abfrage von funktionellen Daten erfolgt dabei über die Bereitstellung von "links" zu weiteren im Internet verfügbaren Protein- oder DNA Datenbanken, wie z.B. Medline oder OMIM (siehe Zusammenfassung, Seite 173, zweite Säule, zweiter Absatz bis Seite 174,zweite Säule, dritter Absatz; Seite 175, Tabelle 1, erste Säule, erster Absatz; Seite 180, Abbildung 5). Der Gegenstand von Anspruch 1 unterscheidet sich davon indem zwei Homologietests durchgeführt werden. Dabei werden nach Durchführung des ersten Tests, alle DNA-Sequenzen, die über einem vorgegebenen Homologieschwellenwert liegen ausgesondert. Dies führt dazu, daß der Rechenaufwand bei der Durchführung des zweiten Sequenzvergleichs reduziert wird.

Aufgabe der vorliegenden Anmeldung war es demnach ein schnelleres DNA-Sequenzvergleich Verfahren zu entwickeln.

Gelöst wurde das oben genannte Problem, indem der Sequenzdatenbestand der zweiten Datenbank zuerst durch Subtraktion mit bereits bekannten Daten aus der ersten Datenbank reduziert wird, so daß der zweite Homologievergleich weniger Rechnerkapazität benötigt. Dies führt zu einem beschleunigten Sequenzdatenabgleich. Dieser Verfahrensschritt war aus den zur Verfügung stehenden Dokumenten nicht bekannt. Darüber hinaus erscheint der dadurch erzielte technische Effekt nicht naheliegend aus den zur Verfügung stehenden Dokumenten ableitbar. Eine erfinderische Tätigkeit kann daher anerkannt werden (Art. 33(3) PCT). Dasselbe gilt für den Gegenstand der davon abhängigen Ansprüche 2 bis 14.

Punkt VI:

Die im Recherchenbericht zitierten Dokumente WO-A-0063687, eingereicht am 14.04.2000, am 26.10.2000 veröffentlicht mit Prioritäts-relevantem Datum vom 15.04.1999 und das Dokument WO-A-0113105, eingereicht am 28.07.2000, am 22.01.2001 veröffentlicht mit Prioritäts-relevantem Datum vom 30.07.1999 könnten möglicherweise relevant für den Gegenstand der vorliegenden Anmeldung sein, falls die beanspruchte Priorität der vorliegenden Ansprüche nicht gültig sein sollte.

Punkt VII:

1. Im Widerspruch zu den Erfordernissen der Regel 5.1 a) ii) PCT werden in der Beschreibung weder der in den Dokumenten D1 und D2 offenbarte einschlägige Stand der Technik noch diese Dokumente angegeben.

Punkt VIII:

1. Die in den Ansprüchen 1 und 2 benutzten Ausdrücke:
   Anspruch 1(d): "über einem vorgegebenen Schwellenwert homolog sind"
   Anspruch 1(f): "einen vorgegebenen Schwellenwert überschreitet"
   Anspruch 1(i): "durch linguistische Analyse von Textdefinitionen"
   Anspruch 2(h): "in einer nach vorgebbaren Kriterien optimierten Anpassung an die jeweils homologen Biosequenzen".

sind vage und unklar und lassen den Leser über die Bedeutung der betreffenden technischen Merkmale im Ungewissen. Dies hat zur Folge, daß die Definition des Gegenstands dieser Ansprüche nicht klar ist (Artikel 6 PCT). Unklar sind im besonderen die technischen Kriterien, wodurch die Homologieschwellenwerte bestimmt werden (80%, 90%, 100% Homologie?). In Schritt 1(d) werden zuerst "homologe" Sequenzen aus der Vergleichsdatenbank entfernt, während in einem zweiten Homologietest wiederum "homologe" Sequenzen gefunden werden.
Darüber hinaus ist der Anspruch 2 völlig unklar, was den beanspruchten Umfang und den Gegenstand dieses Anspruchs betrifft.

2. Darüber hinaus wird in der Beschreibung keine technische Lehre gegeben, wie das "Computerprogramm" von Anspruch 1 programmiert wurde. Es scheint daher in einer für den Fachmann nicht nacharbeitbaren Form offenbart zu sein (Art. 5 PCT).

Patentansprüche

1. Verfahren zum Ermitteln potentiell bedeutsamer DNA- und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen) mit den folgenden Schritten:

a) Ermitteln beliebiger Artsequenzen der interessierenden Spezies mit biologischen bzw. gentechnischen Methoden und Speichern der Artsequenzen in einer ersten Datenbank,

b) Erfassen bekannter DNA-/Nukleinsäuresequenzen einer vorgegebenen Gruppe anderer Arten (Biosequenzen) einschließlich der funktionalen Bedeutung dieser Sequenzen, in einer zweiten Datenbank, in welcher die Biosequenzen und Zusatzinformationen einschließlich der funktionalen Bedeutung einzelner Biosequenzen gespeichert sind,

c) Vergleichen der bereits bekannten Artsequenzen der interessierenden Spezies mit den Biosequenzen der in der zweiten Datenbank gespeicherten, vorgegebenen Gruppe von Biosequenzen in einem Homologietest,

d) Aussondern derjenigen Biosequenzen der vorgegebenen Gruppe, die zu den bekannten Artsequenzen über einem vorgegebenen Schwellenwert homolog sind,

e) Vergleichen der aus der zweiten Datenbank verbleibenden, nicht ausgesonderten Biosequenzen aus der erwähnten Gruppe mit den nach Schritt a ermittelten Artsequenzen in einem zweiten Homologietest,

f) Speichern und/oder Ausgeben derjenigen Artsequenzen als Artsequenzen potentiell erhöhter Bedeutung, deren Homologie mit Biosequenzen aus den aus der erwähnten Gruppe verbliebenen Biosequenzen einen vorgegebenen zweiten Schwellenwert überschreitet, zusammen mit Informationen über die hierzu jeweils homologen Biosequenzen,

g) wobei Schritt e) wahlweise auch vor Schritt c) und ohne vorheriges Aussondern gemäß Schritt d) durchführbar ist und

i) Klassifizieren der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen, d. h. Einordnen (Sortieren) in bestimmte Klassen von Sequenzen durch linguistische Analyse von Textdefinitionen der zu den homologen Biosequenzen gespeicherten Zusatzinformationen.

2. Verfahren nach Anspruch 1, gekennzeichnet durch die folgenden weiteren Schritte:

h) Anpassen der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen in einer nach vorgebbaren Kriterien optimierten Anpassung an die jeweils homologen Biosequenzen und Ausgabe und/oder Speicherung charakteristischer Parameter der optimierten Anpassung, wie zum Beispiel der prozentualen Übereinstimmung, der Länge übereinstimmender Sequenzabschnitte und der optimierten relativen Ausrichtung (Alignment).

3. Verfahren nach einem der Ansprüche 1 bis 2, gekennzeichnet durch den folgenden Schritt:

a. Ergänzen der den potentiell bedeutsamen Artsequenzen zuzuordnenden Eigenschaftsinformationen der jeweils homologen Biosequenzen durch Erfassen von Hinweisen (Links) zu den gemäß Schritt f) erfaßten Biosequenzen in der zweiten Datenbank auf mindestens eine dritte Datenbank und Erfassen der zu den erwähnten Biosequenzen in der dritten Datenbank gespeicherten Informationen.

4. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß die dritte Datenbank eine mindestens in Teilbereichen taxonomisch organisierte Klassifikation bereithält.

5. Verfahren nach Anspruch 4, dadurch gekennzeichnet, daß die dritte Datenbank die MEDLINE Datenbank ist.

6. Verfahren nach Anspruch 4, gekennzeichnet durch Vergleichen der nach taxonomischen Kriterien den jeweiligen Biosequenzen zugeordneten Stichworte mit einer vorgegebenen Liste bzw. Datei von Stichworten und Ausgabe übereinstimmender Stichworte sowie der betreffenden Biosequenzen und der homologen Artsequenzen bzw. jeweils einer Kennung derselben, für die übereinstimmende Stichworte mit der vorgegebenen Liste von Stichworten gefunden wurden.

7. Verfahren nach einem der Ansprüche 1 bis 6, dadurch gekennzeichnet, daß der Vergleich einer vorgegebenen (klassifizierten) Liste von Stichworten mindestens mit den Medical Subject Headings der Medline-Datenbank erfolgt.

8. Verfahren nach einem der Ansprüche 1 bis 3, dadurch gekennzeichnet, daß die dritte Datenbank die UNIGENE Datenbank ist.

9. Verfahren nach Anspruch 8, dadurch gekennzeichnet, daß auf der Basis der EST-Clusterpositionen aus UNIGENE Informationen über entsprechende oder benachbarte Sequenzabschnitte aus GENEMAP und/oder GDB erfaßt werden.

10. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß weitere Datenbanken nach Verknüpfungsgliedern zu den in der dritten Datenbank ermittelten Fundstellen durchsucht werden und Hinzufügen der entsprechenden weiteren Informationen bzw. von Hinweisen auf die weiteren Informationen zu den entsprechenden Artsequenzen erhöhter Bedeutung.

11. Verfahren nach einem der Ansprüche 1 bis 10, dadurch gekennzeichnet, daß mindestens die zweite Datenbank eine öffentlich zugängliche Datenbank ist.

12. Verfahren nach einem der Anspruch 4 bis 11, dadurch gekennzeichnet, daß die weiteren Datenbanken aus der Gruppe ausgewählt werden, die aus den Unigene, genemap und GDB (neu) sowie OMIM-, KEGG- und UMLS-Datenbanken besteht.

13. Verfahren nach einem der Ansprüche 1 bis 12, dadurch gekennzeichnet, daß das Hinzufügen weiterer Informationen zu den gemäß Schritt f ermittelten Artsequenzen in einem Pipelineverfahren erfolgt, wobei die hinzugefügten Informationen in Form von Verknüpfungsgliedern zu den zugeordneten Positionen in weiteren Datenbanken bestehen.

14. Verfahren nach einem der Ansprüche 1 bis 13, dadurch gekennzeichnet, daß die interessierende Spezies die menschliche Spezies ist und daß die zugeordnete Gruppe von Biosequenzen die Biosequenzen von wirbellosen Tieren, Säugetieren, Primaten, Nagetieren und Wirbeltieren, sowie die noch nicht klassifizierten Neueinträge der zweiten Datenbank umfaßt.

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
22. März 2001 (22.03.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/20024 A3

(51) Internationale Patentklassifikation[7]: C12Q 1/68, G06F 19/00

(21) Internationales Aktenzeichen: PCT/EP00/07953

(22) Internationales Anmeldedatum: 16. August 2000 (16.08.2000)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
199 41 606.0 1. September 1999 (01.09.1999) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **MERCK PATENT GMBH** [DE/DE]; Frankfurter Strasse 253, 64293 Darmstadt (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **TOLDO, Luca** [DE/DE]; Konrad-Adenauer-Strasse 1, 69514 Laudenbach (DE). **RIPPMANN, Friedrich** [DE/DE]; Schröderstrasse 79, 69120 Heidelberg (DE).

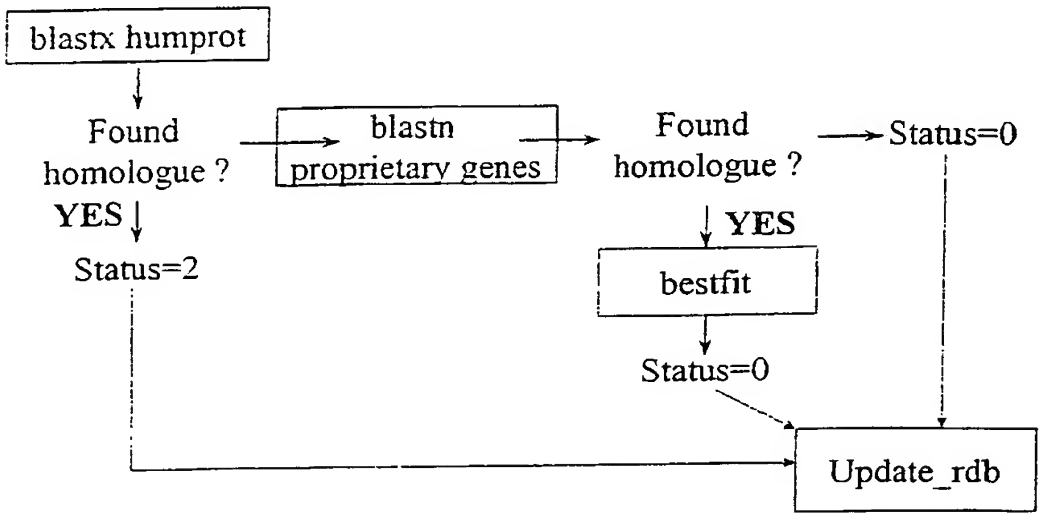
(74) Anwalt: **WEBER - SEIFFERT - LIEKE**; Postfach 61 45, 65051 Wiesbaden (DE).

(81) Bestimmungsstaaten *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW.

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD FOR DETERMINING NUCLEIC AND/OR AMINO ACID SEQUENCES

(54) Bezeichnung: VERFAHREN ZUM ERMITTELN VON NUKLEIN- UND/ODER AMINOSÄURESEQUENZEN



(57) Abstract: The invention relates to a method for determining potentially relevant DNA and/or nucleic acid sequences of a species of interest (species sequences). The aim of the invention is to create a method for determining DNA and/or nucleic acid sequences with which those DNA and/or nucleic acid sequences are specifically selected that have a potentially increased relevance, that is that can be examined with respect to certain functions, especially with regard to a potential relevance for a disease, with a considerably reduced amount of research required.

(57) Zusammenfassung: Die vorliegende Erfindung betrifft ein Verfahren zum Ermitteln potentiell bedeutsamer DNA- und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen). Um ein Verfahren zum Ermitteln von DNA- und/oder Nukleinsäuresequenzen zu schaffen, bei welchem gezielt solche DNA- und/oder Nukleinsäuresequenzen herausselektiert werden, die eine potentiell erhöhte Bedeutsamkeit haben, das heißt die mit erheblich weniger Forschungsaufwand gezielt im Hinblick auf bestimmte Funktionen untersucht werden können, insbesondere im Hinblick auf eine potentielle Kranheitsrelevanz.

WO 01/20024 A3

(84) **Bestimmungsstaaten** *(regional)*: ARIPO-Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI-Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *mit internationalem Recherchenbericht*

(88) **Veröffentlichungsdatum des internationalen Recherchenberichts:** 23. Mai 2002

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

IN NATIONAL SEARCH REPORT

International Application No
PCT/EP 00/07953

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68 G06F19/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	MADDEN T L ET AL: "APPLICATIONS OF NETWORK BLAST SERVER" METHODS IN ENZYMOLOGY, ACADEMIC PRESS INC, SAN DIEGO, CA, US, vol. 266, 1996, pages 131-141, XP001006313 ISSN: 0076-6879 the whole document --- -/--	1-15

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 August 2001	09/08/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Filloy García, E

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WORLEY K C ET AL: "BEAUTY: AN ENHANCED BLAST-BASED SEARCH TOOL THAT INTEGRATES MULTIPLE BIOLOGICAL INFORMATION RESOURCES INTO SEQUENCE SIMILARITY SEARCH RESULTS" GENOME RESEARCH,US,COLD SPRING HARBOR LABORATORY PRESS, vol. 5, no. 2, 1 September 1995 (1995-09-01), pages 173-184, XP000534406 ISSN: 1088-9051 the whole document	1-15
E	WO 01 13105 A (CHIN DANIEL J ;HENDRIX DONNA (US); ZHAO OLIVER (US); AGY THERAPEUT) 22 February 2001 (2001-02-22) abstract; claims 1-13	1-15
E	WO 00 63687 A (UNIV COLUMBIA) 26 October 2000 (2000-10-26) abstract; claim 1 page 44, line 5 -page 45, line 10	1-15
A	US 5 871 697 A (DEEM MICHAEL W ET AL) 16 February 1999 (1999-02-16) abstract; claims 1-6 column 58, paragraph 2 -column 59, paragraph 2	1-15

IN NATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP 00/07953

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0113105 A	22-02-2001	AU 6611900 A	13-03-2001
WO 0063687 A	26-10-2000	AU 4355600 A	02-11-2000
US 5871697 A	16-02-1999	AU 730830 B	15-03-2001
		AU 7476396 A	15-05-1997
		EP 0866877 A	30-09-1998
		JP 2000500647 T	25-01-2000
		WO 9715690 A	01-05-1997
		US 6231812 B	15-05-2001
		US 5972693 A	26-10-1999
		US 2001007985 A	12-07-2001
		US 6141657 A	31-10-2000

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP 00/07953

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 C12Q1/68 G06F19/00

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G06F

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, WPI Data

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
Y	MADDEN T L ET AL: "APPLICATIONS OF NETWORK BLAST SERVER" METHODS IN ENZYMOLOGY,ACADEMIC PRESS INC, SAN DIEGO, CA,US, Bd. 266, 1996, Seiten 131-141, XP001006313 ISSN: 0076-6879 das ganze Dokument --- -/--	1-15

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

° Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
2. August 2001	09/08/2001

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Filloy García, E

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
Y	WORLEY K C ET AL: "BEAUTY: AN ENHANCED BLAST-BASED SEARCH TOOL THAT INTEGRATES MULTIPLE BIOLOGICAL INFORMATION RESOURCES INTO SEQUENCE SIMILARITY SEARCH RESULTS" GENOME RESEARCH,US,COLD SPRING HARBOR LABORATORY PRESS, Bd. 5, Nr. 2, 1. September 1995 (1995-09-01), Seiten 173-184, XP000534406 ISSN: 1088-9051 das ganze Dokument	1-15
E	WO 01 13105 A (CHIN DANIEL J ;HENDRIX DONNA (US); ZHAO OLIVER (US); AGY THERAPEUT) 22. Februar 2001 (2001-02-22) Zusammenfassung; Ansprüche 1-13	1-15
E	WO 00 63687 A (UNIV COLUMBIA) 26. Oktober 2000 (2000-10-26) Zusammenfassung; Anspruch 1 Seite 44, Zeile 5 -Seite 45, Zeile 10	1-15
A	US 5 871 697 A (DEEM MICHAEL W ET AL) 16. Februar 1999 (1999-02-16) Zusammenfassung; Ansprüche 1-6 Spalte 58, Absatz 2 -Spalte 59, Absatz 2	1-15

INTERNATIONAL ECHERCHENBERICHT

Angaben zu Veröffentlic. gen, die zur selben Patentfamilie gehören

ationales Aktenzeichen
PCT/EP 00/07953

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
WO 0113105 A	22-02-2001	AU 6611900 A	13-03-2001
WO 0063687 A	26-10-2000	AU 4355600 A	02-11-2000
US 5871697 A	16-02-1999	AU 730830 B	15-03-2001
		AU 7476396 A	15-05-1997
		EP 0866877 A	30-09-1998
		JP 2000500647 T	25-01-2000
		WO 9715690 A	01-05-1997
		US 6231812 B	15-05-2001
		US 5972693 A	26-10-1999
		US 2001007985 A	12-07-2001
		US 6141657 A	31-10-2000

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
22. März 2001 (22.03.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/20024 A2

(51) Internationale Patentklassifikation[7]: C12Q 1/68

(21) Internationales Aktenzeichen: PCT/EP00/07953

(22) Internationales Anmeldedatum:
16. August 2000 (16.08.2000)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
199 41 606.0 1. September 1999 (01.09.1999) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **MERCK PATENT GMBH** [DE/DE]; Frankfurter Strasse 253, 64293 Darmstadt (DE).

(72) Erfinder; und
(75) Erfinder/Anmelder *(nur für US)*: **TOLDO, Luca** [DE/DE]; Konrad-Adenauer-Strasse 1, 69514 Laudenbach (DE). **RIPPMANN, Friedrich** [DE/DE]; Schröderstrasse 79, 69120 Heidelberg (DE).

(74) Anwalt: **WEBER - SEIFFERT - LIEKE**; Postfach 61 45, 65051 Wiesbaden (DE).

(81) Bestimmungsstaaten *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Bestimmungsstaaten *(regional)*: ARIPO-Patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), eurasisches Patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI-Patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:
— *Ohne internationalen Recherchenbericht und erneut zu veröffentlichen nach Erhalt des Berichts.*

Zur Erklärung der Zweibuchstaben-Codes, und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

(54) Title: METHOD FOR DETERMINING NUCLEIC AND/OR AMINO ACID SEQUENCES

(54) Bezeichnung: VERFAHREN ZUM ERMITTELN VON NUKLEIN- UND/ODER AMINOSÄURESEQUENZEN

(57) Abstract: The invention relates to a method for determining potentially relevant DNA and/or nucleic acid sequences of a species of interest (species sequences). The aim of the invention is to create a method for determining DNA and/or nucleic acid sequences with which those DNA and/or nucleic acid sequences are specifically selected that have a potentially increased relevance, that is that can be examined with respect to certain functions, especially with regard to a potential relevance for a disease, with a considerably reduced amount of research required.

(57) Zusammenfassung: Die vorliegende Erfindung betrifft ein Verfahren zum Ermitteln potentiell bedeutsamer DNA- und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen). Um ein Verfahren zum Ermitteln von DNA- und/oder Nukleinsäuresequenzen zu schaffen, bei welchem gezielt solche DNA- und/oder Nukleinsäuresequenzen herausselektiert werden, die eine potentiell erhöhte Bedeutsamkeit haben, das heißt die mit erheblich weniger Forschungsaufwand gezielt im Hinblick auf bestimmte Funktionen untersucht werden können, insbesondere im Hinblick auf eine potentielle Kranheitsrelevanz.

WO 01/20024 A2

Verfahren zum Ermitteln von Nuklein- und/oder Aminosäuresequenzen

Die vorliegende Erfindung betrifft ein Verfahren zum Erfassen von DNA- und/oder Nukleinsäuresequenzen und insbesondere ein Verfahren zur Erfassung solcher DNA- und/oder Nukleinsäuresequenzen einer gegebenen Spezies (im folgenden kurz als „Artsequenzen" bezeichnet), die eine potentiell erhöhte Bedeutsamkeit haben und die somit besonders lohnenswert erscheinende Forschungsobjekte sind.

Die Biowissenschaften und insbesondere die Gentechnologie haben in den vergangenen Jahren eine besonders rasante Entwicklung durchlaufen. Grundlage hierfür waren zum Beispiel neue Verfahren zur Erzeugung und Vervielfältigung von gentechnischem Material, wie z. B. die Polymerase-Kettenreaktion (PCR) und immer bessere Methoden zur Aufspaltung genetischen Materials und zur Identifizierung der Bruchstücke im Detail, das heißt der genauen Abfolge von Nukleinsäuren, die entlang eines Genabschnittes angeordnet sind.

Dies hat dazu geführt, daß die Zahl der in ihrem genauen Aufbau ermittelten Genabschnitte verschiedener Arten immer schneller angewachsen ist und weiterhin anwächst. Ein sehr anspruchsvolles, aber in wenigen Jahren womöglich bereits erreichtes Ziel liegt in der vollständigen Erfassung des menschlichen Genoms, das heißt der Erfassung sämtlicher Sequenzen, aus denen die menschlichen Gene zusammengesetzt sind, einschließlich der genauen Reihenfolge von Nukleinsäuren innerhalb der Sequenzen und der relativen Anordnung der einzelnen Sequenzen zueinander.

Auch wenn die Anordnung und Positionierung bestimmter Sequenzen schon eine nützliche Zusatzinformation bei der Ermittlung der funktionellen Bedeutung der betreffenden Sequenzen liefern kann, so ist doch die reine Kenntnis einer bestimmten Sequenz (Nukleinsäure- oder DNA-Sequenz) nur von sehr geringem Wert, solange man die genaue Funktion und Bedeutung des betreffenden Genabschnittes nicht erkannt und verstanden hat. Gerade dies spielt aber in der wissenschaftlichen Forschung und insbesondere in der Medizin eine immer größere Rolle. So sind z. B. bestimmte Krankheiten mit der konkreten Ausgestaltung ganz bestimmter Genabschnitte aufs Engste verknüpft und die genaue Kenntnis des funktionellen Zusammenhanges zwischen einem bestimmten Genabschnitt und der Ausprägung eines bestimmten Krankheitsbildes kann daher von enormer therapeutischer Bedeutung sein, da sich dann viel leichter Pharmazeutika entwickeln lassen, die exakt ein krankhaftes Defizit ausgleichen. Gegebenenfalls kann sogar eine vollständige Heilung dadurch herbeigeführt werden, daß durch Gabe eines Therapeutikums, z. B eines Inhibitors eines Genproduktes eines kranheitsrelevanten Gens, der gesunde

Gleichgewichtszustand wieder erreicht wird. Dies gilt selbstverständlich nicht nur für die menschliche Spezies, sondern im Prinzip für jede Art von Lebewesen, das heißt sowohl für alle Tier- und Pflanzenarten als auch für mikrobiologische Arten.

Wie bereits erwähnt, ist das reine Auffinden immer neuer DNA- oder Nukleinsäuresequenzen ohne Kenntnis von deren funktionaler Bedeutung eine relativ nutzlose Sammlung von Daten, da es kaum möglich ist, in gezielter biologischer bzw. medizinischer Forschung die funktionale Bedeutung einzelner Sequenzen oder Sequenzabschnitte auch nur annähernd in dem Tempo zu ermitteln, in dem neue Sequenzen ermittelt werden.

Außerdem ist die Ermittlung der funktionellen Bedeutung von DNA-Sequenzen, auf deren Funktion es keinerlei Hinweise gibt, neben dem dafür erforderlichen Zeitaufwand auch außerordentlich kosten- und personalintensiv und bindet damit viele Kapazitäten.

Ausgehend von diesem Stand der Technik liegt der vorliegenden Erfindung die Aufgabe zugrunde, ein Verfahren zum Ermitteln von DNA- und/oder Nukleinsäuresequenzen zu schaffen, bei welchem gezielt solche DNA- und/oder Nukleinsäuresequenzen herausselektiert werden, die eine potentiell erhöhte Bedeutsamkeit haben, das heißt die mit erheblich weniger Forschungsaufwand gezielt im Hinblick auf bestimmte Funktionen untersucht werden können, insbesondere im Hinblick auf eine potentielle Kranheitsrelevanz, als dies bei den übrigen, nicht auf diese Weise selektierten DNA-Sequenzen möglich wäre.

Diese Aufgabe wird durch die Merkmale des Anspruchs 1 gelöst, wobei die abhängigen Ansprüche vorteilhafte Ausgestaltungen der Erfindung darstellen, durch die die Selektion nochmals verfeinert wird und durch die zusätzliche Informationen gewonnen werden, welche den notwendigen Forschungsaufwand noch weiter reduzieren.

Das erfindungsgemäße Verfahren besteht aus mehreren Schritten, wobei die Reihenfolge der nachstehend aufgelisteten Schritte jedoch mindestens teilweise auch variabel ist. Zum Beispiel könnten zunächst die Schritte b und c und anschließend erst der Schritt a des Anspruchs 1 ausgeführt werden.

Gemäß Schritt a werden im Prinzip beliebige Artsequenzen einer interessierenden Spezies mit biologischen bzw. gentechnischen Methoden ermittelt. Die ermittelten Artsequenzen werden in einer üblichen Nomenklatur als Buchstabencode, der z. B. aus vier Buchstaben besteht, in einer ersten Datenbank gespeichert.

Weiterhin werden gemäß Schritt b alle bekannten DNA- und/oder Nukleinsäuresequenzen einer vorgegebenen Gruppe biologischer Arten oder Klassen in einer zweiten Datenbank erfaßt, in der

im allgemeinen auch die funktionalen Bedeutungen solcher Sequenzen zusammen mit den Sequenzen gespeichert sind. Derartige, öffentlich zugängliche Datenbanken enthalten außerdem mitunter weitere Zusatzinformationen zu den einzelnen Sequenzen. Lediglich zur besseren Unterscheidung werden diese von mehreren Arten stammenden Sequenzen hier kurz als „Biosequenzen" bezeichnet, während Sequenzen der interessierenden Spezies hier durchgehend als „Artsequenzen" bezeichnet werden. Die vorgegebene Gruppe von Arten oder Klassen kann, muß jedoch nicht die interessierende Spezies enthalten. Im Gegenteil, gemäß der vorliegenden Erfindung sind es gerade die über andere Arten in solchen Datenbanken enthalten Informationen, die gemäß der vorliegenden Erfindung mit einem ausgeklügelten Verfahren selektiert werden, die dann durch die Verknüpfung mit anderen Informationsquellen mit Hilfe des erfindungsgemäßen Verfahrens indirekt Hinweise auf die Bedeutung bestimmter Sequenzen der interessierenden Art liefern.

Gemäß Schritt c werden die in einer Datenbank gemäß Schritt b erfaßten Biosequenzen mit den ebenfalls bereits bekannten und möglicherweise in derselben Datenbank gespeicherten Artsequenzen (der interessierenden Art) in einem Homologietest verglichen, wobei wegen der relativ großen Zahl der miteinander zu vergleichenden Sequenzen ein möglichst einfacher Homologietest verwendet werden sollte. Liegt dann die Homologie zwischen den bekannten Artsequenzen und den bekannten Biosequenzen über einem gewissen Schwellenwert, so werden all diese zu bekannten Artsequenzen homologen Biosequenzen gemäß Schritt d aus dem weiter zu betrachtenden Datenbestand ausgesondert. Damit ist die Menge der verbleibenden, bekannten Biosequenzen gegenüber den insgesamt öffentlich bekannten Biosequenzen nicht nur durch eine Beschränkung auf eine Gruppe bestimmter Arten reduziert, sondern darüber hinaus auch noch auf diejenigen Sequenzen, zu denen bisher keine homologen Artsequenzen ermittelt wurden.

Die gemäß Schritt a gespeicherten bzw. neu ermittelten DNA-/Nukleinsäuresequenzen werden dann in Schritt e mit diesem verbleibenden, reduzierten Bestand an Biosequenzen in einem Homologietest verglichen. Zweckmäßigerweise werden zur Bestätigung der Homologie und zum besseren Verständnis der übereinstimmenden Abschnitte der Sequenzen die Artsequenz und die hierzu homologe Biosequenz aneinander angepaßt. Wenn die Homologie über einem vorgegebenen zweiten Grenzwert liegt, so werden die betreffenden Biosequenzen gemäß Schritt f zusammen mit mindestens einem die zugehörige Biosequenz eindeutig identifizierenden Verknüpfungsglied abgespeichert bzw. als potentiell bedeutsame Artsequenz ausgegeben.

Durch die Verknüpfung mit einer oder mehreren bestimmten Biosequenzen, zu denen bereits Funktionsbeschreibungen und andere Zusatzinformationen bekannt sind, kann man sehr gezielt nach analogen Funktionen der neu ermittelten Artsequenzen suchen und hat dabei auch einen sehr hohen Grad von Erfolgswahrscheinlichkeit mit verhältnismäßig niedrigem Aufwand. Diese erhöhte Erfolgswahrscheinlichkeit bei niedrigem Aufwand macht die betreffenden Artsequenzen

zu Artsequenzen potentiell erhöhter Bedeutsamkeit, da andere von ihrer Struktur und Länge her gleichwertige Artsequenzen, zu denen aber keine Homologe mit bekannten Funktionen existieren, einen erheblich größeren Aufwand bei der Ermittlung ihrer funktionellen Bedeutung erfordern würden.

Allgemein werden verschiedene Informationspools durch die vorliegende Erfindung auf eine besondere, strategisch günstige Weise so miteinander verknüpft, daß ein Maximum an Information zu einer Sequenz mit einem in der Praxis noch machbaren Minimum an Aufwand gewonnen wird. Dagegen würde eine nach üblichen mathematischen Kriterien vorgenommene wechselseitige Verknüpfung aller zu jeweils einer Sequenz und deren Homologen gespeicherten Daten aus einer größeren Gruppe von biomedizinischen Datenbanken, wie sie vorliegend verwendet werden, alle derzeit verfügbaren Rechenkapazitäten bei weitem übersteigen.

Bei dem erfindungsgemäßen Verfahren lassen sich daher nicht nur wesentlich schneller und sicherer Erfolge bei der Entwicklung von Medikamenten und der Therapie von Krankheiten erzielen, sondern es ist diese Erfolgswahrscheinlichkeit bei gleichzeitig reduziertem Forschungsaufwand beträchtlich erhöht.

Um diesen Aufwand noch weiter zu reduzieren, ist in einer bevorzugten Ausführungsform der Erfindung vorgesehen, daß gemäß einem weiteren Schritt g in den öffentlich zugänglichen Datenbanken Hinweise (Links) erfaßt werden, die dort zu Biosequenzen in der zweiten, öffentlichen Datenbank gespeichert sind, und zwar zu den Biosequenzen, die zuvor als Homologe zu neuen Artsequenzen ermittelt wurden, wobei vorzugsweise solche Hinweise ausgewertet und verwendet werden, die auf eine taxonomisch organisierte Datenbank hinweisen. Eine solche taxonomisch organisierte Datenbank enthält zu den jeweiligen Biosequenzen nach einheitlichen wissenschaftlichen Kriterien ausgewählte Stichworte, die dann gemäß Schritt h mit einer vorgegebenen Liste von Stichworten verglichen werden, wobei diese Liste wiederum so ausgewählt ist, daß sie die Forschungsgebiete eines Benutzers abdeckt. Die betreffende Biosequenz und die zugehörige Artsequenz werden also nur dann in dem als lohnende Zielobjekte zu definierenden Datenbestand erhalten, wenn Übereinstimmungen zwischen einer vorgegebenen Stichwortliste und den nach taxonomischen Kriterien vergebenen Stichwörtern in der entsprechenden Datenbank (dritte Datenbank) bestehen. Die betreffenden Stichwörter, die in gewisser Weise funktionale Bedeutungen repräsentieren, lassen dann wiederum eine gezieltere Forschung nach den speziellen Eigenschaften einer Artsequenz zu.

Die Datenbank, in welcher neu ermittelte Artsequenzen für eine weitere Untersuchung gespeichert werden, kann eine öffentliche Datenbank sein, dürfte im Regelfall aber eine private Datenbank sein, zu der jeweils nur der Benutzer oder einige wenige Benutzer Zugang haben, jedoch nicht die Öffentlichkeit.

Dagegen hat die zweite Datenbank, in der auch Zusatzinformationen zu den betreffenden Biosequenzen und Hinweise auf andere Datenbanken und darin gespeicherte Informationen enthalten sind, im allgemeinen eine öffentliche Zugangsmöglichkeit.

Eine für die Zwecke der vorliegenden Erfindung besonders geeignete dritte Datenbank, die nach taxonomischen Kriterien ausgewählte Stichworte (MeSH Begriffe) enthält, ist die sogenannte „MEDLINE"-Datenbank. Diese Datenbank enthält zum einen eine Identifikationsnummer für jede biomedizinische Literaturstelle und zusätzliche Informationen zusammen mit einer Reihe weiterer Daten, und unter anderem auch Stichworte, die als „medical subject headings" bezeichnet werden. Darüber hinaus gibt es Hinweise auf Fundstellen, Autoren, Veröffentlichungen. Und sogenannte RN Nummern.

Daneben enthält die MEDLINE Datenbank einen sogenannten Sequenz Identifier, der vorzugsweise als eines der notwendigen Verknüpfungsglieder benutzt wird.

Auf diese Weise ist es möglich, für einen Benutzer umfassende Informationen zu erzeugen und zusammenzustellen, der ursprünglich lediglich DNA-/Nukleinsäuresequenzen vorliegen hatte, zu denen keinerlei Informationen bekannt waren, wobei durch das erfindungsgemäße Verfahren automatisch auf dem Weg über Homologietests und das gezielte Filtern und Aussondern von Informationsquellen umfassende Informationen zu einer Artsequenz erzeugt werden, die Bedeutung und Funktion der Sequenz charakterisieren und eine gezielte Forschung ermöglichen. Alle Artsequenzen, für die auf diese Weise Funktionen und Bedeutungen ermittelt werden können, werden um diese Zusatzinformationen ergänzt. Sie können jedoch jederzeit wieder aufgegriffen werden, wenn der Datenbestand in der zweiten (öffentlich zugänglichen) Datenbank entsprechend erweitert worden ist, so daß sich auf diese Weise auch zunächst ausgesonderte Artsequenzen bei einem späteren Durchlauf als lohnenswerte Zielobjekte herausstellen können.

Die Homologietests, die zwischen Artsequenzen und Biosequenzen durchgeführt werden, werden vorzugsweise in einem Pipelineverfahren durchgeführt, so daß nicht immer komplette Datenbestände erfaßt und verwaltet werden müssen.

Weiterhin ist es zweckmäßig, wenn auch über die bereits erwähnten Datenbanken hinaus weitere Datenbanken nach Verknüpfungen insbesondere mit der dritten Datenbank (MEDLINE) durchsucht werden, um im Falle einer entsprechenden Verknüpfung auch die Zusatzinformationen aus diesen zusätzlichen Datenbanken zu verwerten. Hierzu zählen insbesondere auch die als „OMIM" und „KEGG" bezeichneten Datenbanken.

Auch ohne weitere Ausführungen wird davon ausgegangen, daß ein Fachmann die obige Beschreibung im weitesten Umfang nutzen kann. Die bevorzugten Ausführungsformen und Beispiele sind deswegen lediglich als beschreibende, keineswegs als in irgendeiner Weise limitierende Offenbarung aufzufassen.

Die vollständige Offenbarung aller vor- und nachstehend aufgeführten Anmeldungen, Patente und Veröffentlichungen, sowie der korrespondierenden Anmeldung 199 41 606.0, eingereicht am 1. September 1999 sind durch Bezugnahme in diese Anmeldung eingeführt.

Ein Ausführungsbeispiel der Erfindung wird im folgenden anhand von Figuren erläutert, woraus sich weitere Vorteile, Merkmale und Anwendungsmöglichkeiten der vorliegenden Erfindung ergeben. Es zeigen:

Fig. 1 ein Schema zur Reduktion der ermittelten Artsequenzen, wie es den Schritten a bis f in Anspruch 1 entspricht,

Fig. 2 ein Schema von Datenbanken und Datenbankverknüpfungen, wie sie für das weitere Auswerten von Informationen gemäß der vorliegenden Erfindung verwendet werden und

Fig. 3 die Wiedergabe einer Bildschirmdarstellung mit Bedienfeldern und Informationsfeldern zur einer (hypothetischen) Nukleinsäuresequenz.

Generell werden zunächst alle z. B. im Laufe einer Woche neu ermittelten DNA-Sequenzen bzw. Nukleinsäuresequenzen in einer üblichen Nomenklatur (in den standardmäßigen Buchstabencodes) in einer Datenbank gespeichert, wobei außerdem noch eine Identifikationsnummer oder irgendeine andere Codierung zur Identifikation der betreffenden Sequenz vergeben und gleichzeitig abgespeichert wird. Weitere, zusätzlich mit abzuspeichernde Informationen sind z. B. die Sequenzlänge, die Art und andere Zusatzinformationen, die unmittelbar zusammen mit der Ermittlung einer solchen Sequenz zur Verfügung stehen. Die folgenden Verfahrensschritte laufen dann automatisch ab. Es wird auf eine öffentlich zugängliche Sequenzdatenbank zugegriffen, die DNA- und/oder Nukleinsäuresequenzen der verschiedenen Arten enthält. Dabei wird durch die ursprüngliche Eingabe der interessierenden Spezies (z. B. *Homo sapiens*) bereits eine Einschränkung auf eine bestimmte Gruppe von Arten vorgenommen, von denen man sinnvollerweise eine Korrelation und funktionale Ähnlichkeit zu Genabschnitten der interessierenden Art vermuten kann.

Die öffentliche Sequenzdatenbank enthält bereits Daten über die interessierende Art. Daher wird zunächst ein Homologietest zwischen den in der öffentlichen Datenbank dokumentierten Sequenzen der interessierenden Art mit den Biosequenzen der entsprechend ausgewählten Gruppe

von Arten, die in derselben Datenbank gespeichert sind. Dabei werden alle Biosequenzen, die homolog zu den bereits in der öffentlichen Datenbank gespeicherten Artsequenzen sind ausgesondert, da sie offenbar schon Gegenstand entsprechender Forschungen waren bzw. sind.

Zweckmäßigerweise werden die Ergebnisse dieses Verfahrensschrittes protokolliert, so daß bei einer Wiederholung desselben Vorganges z. B. eine Woche später alle bereits einmal ausgesonderten Biosequenzen von vornherein außer Betracht bleiben, was den Verfahrensablauf beträchtlich beschleunigt. Der Homologietest kann sich dann auf die neu hinzugekommenen Biosequenzen beschränken bzw. umgekehrt die zuvor nicht ausgesonderten Biosequenzen müssen noch in einem Homologietest mit neu hinzugekommenen Artsequenzen verglichen werden.

Damit wird jedoch der Ausgangsdatenbestand beträchtlich verringert.

Die noch verbleibenden Biosequenzen werden dann mit den neu ermittelten Artsequenzen in einem Homologietest verglichen. Dabei werden im Regelfall für einige der neu ermittelten Artsequenzen homologe Biosequenzen gefunden. Sodann wird eine Liste bzw. Tabelle der Artsequenzen und der dazu neu gefundenen, homologen Biosequenzen angefertigt und in diese Tabelle bzw. Liste werden auch zusätzliche Informationen aus der öffentlichen Datenbank übernommen, wie z. B. eine medline-Identitätsnummer, die möglicherweise zu einer bekannten Biosequenz gespeichert ist.

Ein weiterer Schritt (h) des Verfahrens besteht im Klassifizieren der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen, d. h. Einordnen (Sortieren) in bestimmte Klassen von Sequenzen durch linguistische Analyse von Textdefinitionen der zu den homologen Biosequenzen gespeicherten Zusatzinformationen. Dies ermöglicht eine Aufteilung in Teildatensätze, die für deren Ergänzung wiederum nur ein Teil der sonstigen Datenbasen in Frage kommt

Weiterhin erfolgt gemäß Schritt i ein Ergänzen der den potentiell bedeutsamen Artsequenzen zuzuordnenden Eigenschaftsinformationen der jeweils homologen Biosequenzen durch Erfassen von Hinweisen (Links) zu den gemäß Schritt f) erfaßten Biosequenzen in der zweiten Datenbank auf mindestens eine dritte Datenbank und Erfassen der zu den erwähnten Biosequenzen in der dritten Datenbank gespeicherten Informationen

Die dritte Datenbank sollte eine mindestens in Teilbereichen taxonomisch organisierte Klassifikation bereitstellen, vorzugsweise handelt es sich dabei um die sogenannte MEDLINE Datenbank.

Erfindungsgemäß werden die nach taxonomischen Kriterien den jeweiligen Biosequenzen zugeordneten Stichworte mit einer vorgegebenen Liste bzw. Datei von Stichworten verglichen und übereinstimmende Stichworte sowie die betreffenden Biosequenzen und die homologen Artse-

quenzen bzw. jeweils eine Kennung derselben, für die übereinstimmende Stichworte mit der vorgegebenen Liste von Stichworten gefunden wurden, werden ausgegeben.

Neben der MEDLINE Datenbank oder auch ersatzweise hierfür werden auch Informationen aus weiteren Datenbanken verwendet, die z. B. aus der Gruppe ausgewählt werden, die aus den Unigene, Genemap und GDB (neu) sowie OMIM-, KEGG- und UMLS-Datenbanken besteht.

In erster Linie ist die interessierende Spezies die des *Homo sapiens*, wobei aber das erfindungsgemäße Verfahren für eine andere Spezies mit im wesentlichen ähnlicher Zielsetzung ebenso verwendet werden kann.

Mit Bezug auf die Figuren werden nun der Ablauf und das Ergebnis eines hypothetischen Ausführungsbeispiel etwas genauer erläutert. Wie bereits erwähnt, werden gemäß Schritt c in Patentanspruch 1 bereits bekannte Artsequenzen der interessierenden Spezies mit den Biosequenzen in einem Homologietest verglichen, die zu einer vorgegebenen Gruppe von Biosequenzen gehören, welche in der zweiten Datenbank gespeichert sind. Dieser Schritt ist in Fig. 1 mit "blastx humprot" bezeichnet. Sofern homologe Sequenzen gefunden wurden, wird den zu den bereits bekannten Artsequenzen homologen Biosequenzen ein bestimmter Status (hier Status = 2) zugeordnet und diese Biosequenzen werden entsprechend gekennzeichnet und aus dem interessierenden Pool der zweiten Datenbank ausgesondert.

Anschließend erfolgt mit den Artsequenzen, die gemäß Schritt a ermittelt wurden, ein weiterer Homologietest mit den aus der zweiten Datenbank verbleibenden Biosequenzen, die bis dahin noch nicht als Homologe zu bekannten Artsequenzen ermittelt wurden. Dieser Schritt ist in Fig. 1 mit "Blastn proprietary genes" bezeichnet. Sofern homologe Biosequenzen gefunden wurden, erfolgt die bestmögliche Anpassung und Ausrichtung (dieser Schritt ist in Fig. 1 mit "bestfit" bezeichnet) und die die Anpassung, Länge und Ausrichtung kennzeichnenden Daten werden zusammen mit der betreffenden Sequenz gespeichert. Der den entsprechenden Biosequenzen zugeordnete Status 0 bedeutet, daß diese Biosequenzen weiterhin in dem interessierenden Pool an Daten verbleiben.

Ebenso verbleiben auch diejenigen Biosequenzen in dem interessierenden und reduzierten Datenpool, zu welchen weder unter den ermittelten Artsequenzen noch unter den bereits bekannten Artsequenzen Homologe zu finden waren.

Auf diese Weise werden Datensätze erzeugt, welchen neu ermittelten Artsequenzen entsprechende homologe Biosequenzen zugeordnet sind. Der Benutzer des erfindungsgemäßen Systems bedient dieses zweckmäßigerweise von einem Bildschirmarbeitsplatz mit entsprechenden Einrichtungen. In Fig. 3 ist schematisch eine Bildschirmanzeige wiedergegeben, die ein hypothe-

tisches Ergebnis einer Ermittlung potentiell bedeutsamer Artsequenzen gemäß der Erfindung zeigt. Dabei ist allerdings darauf hinzuweisen, daß das dargestellte Ergebnis kein Realerzeugnis, sondern lediglich ein hypothetisches, künstlich synthetisiertes Ergebnis ist, an welchem jedoch prinzipiell alle wesentlichen Schritte und Ergebnisse eines typischen Ausführungsbeispiels abgelesen werden können.

Der Bildschirm zeigt am linken Rand eine Reihe von Befehls- und Parameterfeldern, die der Benutzer bedienen kann. Beispielsweise wählt er in dem Feld 1.2 einen Grenzwertparameter aus, der die minimale Länge der Homologie zwischen Artsequenz und Biosequenz angibt, die gemäß Homologietest und bestmöglicher Anpassung mit den Nukleinsäuren der homologen Sequenz übereinstimmen. In Feld 1.3 wird der Grenzwert einer prozentualen Übereinstimmung wiedergegeben. In Feld 1.4 kann z.B. ein Stichwort eingegeben werden, welches in Verbindung mit den entsprechenden homologen Sequenzen gesucht werden soll.

Die übrigen Bedienfelder sind selbsterklärend.

Nachdem der/die Benutzer/in entsprechende Parameter ausgewählt hat und das zugrunde liegende Programm startet, erhält er/sie nach kurzer Zeit eine Liste von Artsequenzen, die eine oder mehrere Biosequenzen Homologe haben, welche den Kriterien der Benutzereingabe entsprechen. Zum Beispiel zeigt Abb. 3, daß 124 Artsequenzen eine oder mehrere Biosequenzen haben, welche homolog mit einer prozentualen Identität größer als 95% sind und über eine Homologielänge größer als 500 Basenpaaren verfügen. Darüber hinaus haben die Einträge MeSH Begriffe, die hauptsächlich mit CNS (Zentrales Nerven System) assoziiert sind. Von den 124 Einträgen zeigt Abb. 3 die fünfte Artsequenz, welche mit der Ziffernfolge 44567 bezeichnet ist. Die Biosequenzen, die homolog sind mit der Artsequenz, sind in der rechten Bildhälfte unter "seeds" angegeben. Dabei sind, um diese Zuordnung einzelner Daten aus umfangreichen Dateien zu einer bestimmten vorgegebenen Artsequenz einschließlich der vielen Zusatzinformationen erzeugen zu können, mehrere Schritte notwendig, die allerdings in einem entsprechenden Programm automatisch ablaufen, wobei die Abläufe schematisch an Fig. 2 erläutert werden sollen. Aus dem Homologietest, der in Fig. 1 mit "blast proprietary genes" bezeichnet ist und aus den sich daraus ergebenden Homologen in der zweiten Datenbank, lassen sich aus der zweiten Datenbank sogenannte Genbank Identifier (Genbank ID) ermitteln, die wiederum auch in anderen Datenbanken abgelegt sind, und so eine Relation zwischen verschiedenen Nuklein- und/oder Aminosäuresequenzen und anderen, in den Datenbanken gespeicherten Informationen herstellen.

Eine Schlüsselfunktion kommt dabei der Medline-Datenbank und dem darin festgelegten MEDLINE-Identifier (Block "Medline ID") zu, der in vielen anderen Datenbanken registriert ist. Die unter "seeds" angegebenen Sequenzen sind durch einen Genbank Identifier charakterisiert. Diese

durch den Genbank Identifier bezeichneten Einträge können unter anderem auch Medline Identifier enthalten. Aus der MEDLINE Datenbank lassen sich die Titel der entsprechenden Einträge mit Hilfe dieser Medline identifier ermitteln. Außerdem sind in dieser Datenbank oft auch Hinweise auf bestimmte Enzyme abgelegt, die mit dem betreffenden Genabschnitt in Verbindung gebracht werden und hieraus ergeben sich wiederum die biochemischen Reaktionspfade, die von diesen Enzymen beeinflußt werden. Über den MEDLINE-Identifier lassen sich außerdem weitere Informationen aus anderen Datenbanken gewinnen, z.B. über pathologische Informationen, die Lokalisierung von Genen auf bestimmten Chromosomenabschnitten etc.

Auf dem Bildschirm wird dann nach dem Durchlauf eines entsprechenden Programms eine ganze Reihe von Informationen wiedergegeben, die neben der wahrscheinlichen Lokalisierung der neu ermittelten Artsequenz eine ganze Reihe von Hinweisen auf dessen Funktion, Organverteilung und Krankheitsrelevanz gibt. Im vorliegenden Fall, der, wie bereits erwähnt, nur hypothetische Informationen zu einer Artsequenz wiedergibt, erkennt man beispielsweise neben der Sequenz 44567 die biochemische Bezeichnung, das Erstellungsdatum der Information, bei 17q23 die Position des Genabschnittes auf einem Chromosom. Darunter sind Gene angegeben, die auf demselben Chromosomenarm lokalisiert sind. Aus der UNIGENE-Datenbank stammen Informationen über Cluster aus Genbruchstücken (EST-Cluster), die über eine bestimmte Nummer (Hs.198237) identifiziert werden. Die Anzahl der ESTs in diesem Cluster im Verhältnis zur Gesamtzahl der Komponenten der vorliegenden Sequenz ist mit 54/82 angegeben. Proangiotensin-Angiotensin gibt die wahrscheinlichsten Stoffwechselpfade oder chemischen Reaktionen an, zu welchen die bekannten Biosequenzen gehören. Weiterhin ist mit BRAIN dasjenige Organ angegeben, in welchem die betreffenden Sequenzen am häufigsten gefunden werden. Die Organverteilung der EST-Komponenten wird durch unterschiedliche Balkenlängen veranschaulicht. Der wahrscheinlichste Bereich einer Krankheitsindikation, die in Verbindung mit dem Datenabgleich ermittelt wurde, ist mit CNS angegeben. In der linken Hälfte erkennt man noch eine horizontale Balkenreihe, wobei die Länge dieser Balken jeweils Übereinstimmungen zwischen der Artsequenz und den in der entsprechenden Zeile angegebenen zugehörigen Biosequenzen oder Sequenzabschnitten angegeben wird. Daneben sind die Biosequenzen unter "seeds" im einzelnen aufgelistet, einschließlich ihrer prozentualen Übereinstimmung und der Länge der übereinstimmenden Sequenzabschnitte. Weiterhin sind angegeben die Titel entsprechender Zeitschriften, die Enzyme, und verschiedene Stichworte.

In dem vorliegenden Beispiel wurden durch die erfindungsgemäße Verknüpfung über verschiedene Identifier, Stichwortsuche und taxonomische Auswertung von Datenbanken gewonnene Information aus den meisten der in Fig. 3 angegebenen Datenbanken ermittelt, mit Ausnahme der mit UMLS, SNOMED und ICD9-CM bezeichneten Blöcke. Zur Speicherung der aus dem Verfahren gewonnenen Informationen wird das Knowledge Interchange Format (KIF) verwendet. Dieses Format kann von verschiedenen Knowledge Engineering Werkzeugen wie z.B. Ontolin-

gua verwendet werden, um unter anderem HTML oder XML Dateien zu generieren und weiterführende Methoden der künstlichen Intelligenz (KI) anzuwenden.

Patentansprüche

1. Verfahren zum Ermitteln potentiell bedeutsamer DNA- und/oder Nukleinsäuresequenzen einer interessierenden Spezies (Artsequenzen) mit den folgenden Schritten:

a) Ermitteln beliebiger Artsequenzen der interessierenden Spezies mit biologischen bzw. gentechnischen Methoden und Speichern der Artsequenzen in einer ersten Datenbank,

b) Erfassen bekannter DNA-/Nukleinsäuresequenzen einer vorgegebenen Gruppe anderer Arten (Biosequenzen) einschließlich der funktionalen Bedeutung dieser Sequenzen, in einer zweiten Datenbank, in welcher die Biosequenzen und Zusatzinformationen einschließlich der funktionalen Bedeutung einzelner Biosequenzen gespeichert sind,

c) Vergleichen der bereits bekannten Artsequenzen der interessierenden Spezies mit den Biosequenzen der in der zweiten Datenbank gespeicherten, vorgegebenen Gruppe von Biosequenzen in einem Homologietest,

d) Aussondern derjenigen Biosequenzen der vorgegebenen Gruppe, die zu den bekannten Artsequenzen über einem vorgegebenen Schwellenwert homolog sind,

e) Vergleichen der aus der zweiten Datenbank verbleibenden, nicht ausgesonderten Biosequenzen aus der erwähnten Gruppe mit den nach Schritt a ermittelten Artsequenzen in einem zweiten Homologietest,

f) Speichern und/oder Ausgeben derjenigen Artsequenzen als Artsequenzen potentiell erhöhter Bedeutung, deren Homologie mit Biosequenzen aus den aus der erwähnten Gruppe verbliebenen Biosequenzen einen vorgegebenen zweiten Schwellenwert überschreitet, zusammen mit Informationen über die hierzu jeweils homologen Biosequenzen.

g) Wobei Schritt e) wahlweise auch vor Schritt c) und ohne vorheriges Aussondern gemäß Schritt d) durchführbar ist.

2. Verfahren nach Anspruch 1, gekennzeichnet durch die folgenden weiteren Schritte:

h) Anpassen der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen in einer nach vorgebbaren Kriterien optimierten Anpassung an die jeweils homologen Biosequenzen und Ausgabe und/oder Speicherung charakteristischer Parameter der optimierten Anpas-

sung, wie zum Beispiel der prozentualen Übereinstimmung, der Länge übereinstimmender Sequenzabschnitte und der optimierten relativen Ausrichtung (Alignment).

3. Verfahren nach Anspruch 1, gekennzeichnet durch die folgenden weiteren Schritte:

i) Klassifizieren der in Schritt f) ausgegebenen bzw. gespeicherten Artsequenzen, d. h. Einordnen (Sortieren) in bestimmte Klassen von Sequenzen durch linguistische Analyse von Textdefinitionen der zu den homologen Biosequenzen gespeicherten Zusatzinformationen.

4. Verfahren nach einem der Ansprüche 1 bis 3, gekennzeichnet durch den folgenden Schritt:

k) Ergänzen der den potentiell bedeutsamen Artsequenzen zuzuordnenden Eigenschaftsinformationen der jeweils homologen Biosequenzen durch Erfassen von Hinweisen (Links) zu den gemäß Schritt f) erfaßten Biosequenzen in der zweiten Datenbank auf mindestens eine dritte Datenbank und Erfassen der zu den erwähnten Biosequenzen in der dritten Datenbank gespeicherten Informationen.

5. Verfahren nach einem der Ansprüche 1 bis 4, dadurch gekennzeichnet, daß die dritte Datenbank eine mindestens in Teilbereichen taxonomisch organisierte Klassifikation bereithält.

6. Verfahren nach Anspruch 5, dadurch gekennzeichnet, daß die dritte Datenbank die MEDLINE Datenbank ist.

7. Verfahren nach Anspruch 5, gekennzeichnet durch Vergleichen der nach taxonomischen Kriterien den jeweiligen Biosequenzen zugeordneten Stichworte mit einer vorgegebenen Liste bzw. Datei von Stichworten und Ausgabe übereinstimmender Stichworte sowie der betreffenden Biosequenzen und der homologen Artsequenzen bzw. jeweils einer Kennung derselben, für die übereinstimmende Stichworte mit der vorgegebenen Liste von Stichworten gefunden wurden.

8. Verfahren nach Anspruch 2 und einem der auf Anspruch 2 rückbezogenen Ansprüche, dadurch gekennzeichnet, daß der Vergleich einer vorgegebenen (klassifizierten) Liste von Stichworten mindestens mit den Medical Subject Headings der Medline-Datenbank erfolgt.

9. Verfahren nach einem der Ansprüche 1 bis 4, dadurch gekennzeichnet, daß die dritte Datenbank die UNIGENE Datenbank ist.

10. Verfahren nach Anspruch 9, dadurch gekennzeichnet, daß auf der Basis der EST-Clusterpositionen aus UNIGENE Informationen über entsprechende oder benachbarte Sequenzabschnitte aus GENEMAP und/oder GDB erfaßt werden.

11. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß weitere Datenbanken nach Verknüpfungsgliedern zu den in der dritten Datenbank ermittelten Fundstellen durchsucht werden und Hinzufügen der entsprechenden weiteren Informationen bzw. von Hinweisen auf die weiteren Informationen zu den entsprechenden Artsequenzen erhöhter Bedeutung.

12. Verfahren nach einem der Ansprüche 1 bis 11, dadurch gekennzeichnet, daß mindestens die zweite Datenbank eine öffentlich zugängliche Datenbank .ist.

13. Verfahren nach einem der Anspruch 5 bis 12, dadurch gekennzeichnet, daß die weiteren Datenbanken aus der Gruppe ausgewählt werden, die aus den Unigene, genemap und GDB (neu) sowie OMIM-, KEGG- und UMLS-Datenbanken besteht.

14. Verfahren nach einem der Ansprüche 1 bis 13, dadurch gekennzeichnet, daß das Hinzufügen weiterer Informationen zu den gemäß Schritt f ermittelten Artsequenzen in einem Pipelineverfahren erfolgt, wobei die hinzugefügten Informationen in Form von Verknüpfungsgliedern zu den zugeordneten Positionen in weiteren Datenbanken bestehen.

15. Verfahren nach einem der Ansprüche 1 bis 14, dadurch gekennzeichnet, daß die interessierende Spezies die menschliche Spezies ist und daß die zugeordnete Gruppe von Biosequenzen die Biosequenzen von wirbellosen Tieren, Säugetieren, Primaten, Nagetieren und Wirbeltieren, sowie die noch nicht klassifizierten Neueinträge der zweiten Datenbank umfaßt.

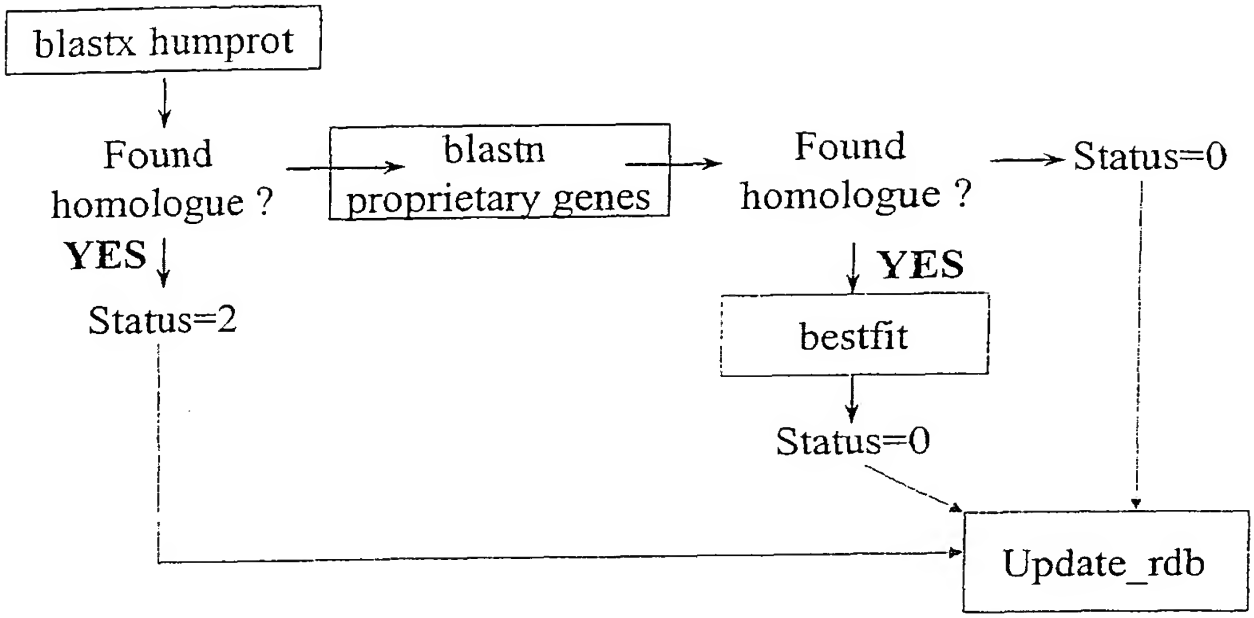

blastx humprot
Found homologue ?
YES
Status=2
blastn proprietary genes
Found homologue ?
YES
bestfit
Status=0
Status=0
Update_rdb

Fig. 1

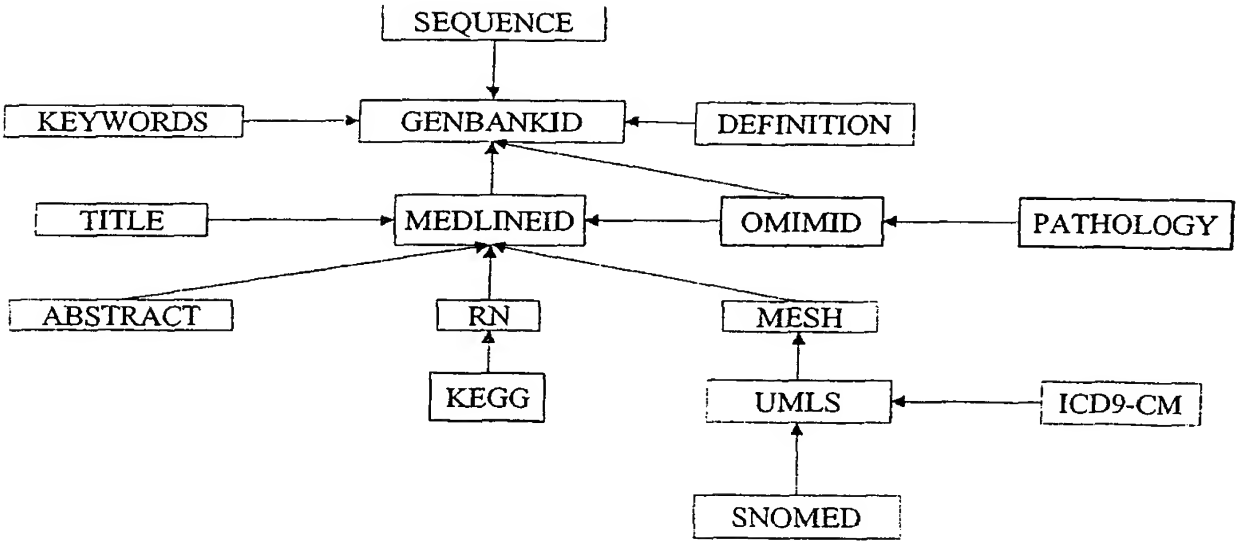

SEQUENCE
KEYWORDS
GENBANKID
DEFINITION
TITLE
MEDLINEID
OMIMID
PATHOLOGY
ABSTRACT
RN
MESH
KEGG
UMLS
ICD9-CM
SNOMED

Fig. 2

5/124

general

>500

>95%

Cns

< > << >>

Keep it

Mail

Save

Print

Exit

Help

Docs

X00244235

44567 **peptidyl-dipeptidase A** **01/03/1997**

17q23 **Proangiotensin-angiotensin** **BRAIN** **CNS**

DCP1
ACE1
CD79B, IGB, B29
PECAM1
TBX2
PRKARIA, TSSI
SMARCD2, BAF60B
DDX5, HLR1, G17P1
ICAM2
UMPH2
APOH
PEPE
MPO
ZNF147, EFP
SCN4A, HYPP, NACIA

Hs.198237

54/82 ANGIOTEN

STARCH

BRAIN

BREAST

BLOOD

BONE

HEART

CNS

DISORDERS

Alzheimer's Disease

SEEDS

RNU03734	peptidyl-dipeptidase A	(84% / 1305 bp)
AA162058	Angiotensin converting enzyme	(83% / 1305 bp)
U56966	Coded for by C.elegans	(80% / 1300 bp)
Z38061	mal5, sta 1, glucoamylase S1	(22% / 850 bp)

TITLES

The isolation of angiotensin-converting enzyme cDNA.

Mouse angiotensin-converting enzyme is a protein composed of two homologous domains

Angiotensin converting enzyme and genetic hypertension: cloning of rat cDNAs and characterization of the enzyme

2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans.

ENZYMES

EC 3.4.15.1

EC 3.2.1.3

TERMS

kidney

Cerebrovascular Disorders

Hypertension

Peptidyl-Dipeptidase A

Fig. 3